A Common Dialect for Infrastructure and Services in Translator

Proposal for a Translator Standards and Reference Implementations Component (SRI)

Principal Investigators:
Melissa Haendel (Oregon State University)
Christopher Mungall (Lawrence Berkeley National Laboratory)
Christopher Bizon (Renaissance Computing Institute, University of North Carolina)

November 15, 2019

Summary Vision Statement

1) We will build a robust, dynamic Translator Standards and Reference Implementations Component (SRI) that integrates the collaborations and investments that the NCATS Translator has made to date. This component will consist of a suite of standards and products, a model for their governance, and processes to coordinate integration and shared implementation:

- **Community governance coordination** will be developed with community buy-in to ensure an effective collaborative environment, and drive consortium-wide consensus on the other components.
- **Architecture and API specifications** will drive community efforts to define details of project architecture and communication protocols across Translator Knowledge Providers (KPs), Autonomous Relay Agents (ARAs), and the Autonomous Relay System (ARS).
- The **BioLink model** will define the standard entity types, relationship types, and a schema shared by all Translator components. This includes related utility libraries and a novel approach to accommodate multiple alternate data modeling perspectives.
- **Integrated reference ontologies** will provide BioLink-compliant terms and relationships. We will draw on the ROBOKOP Ubergraph framework [1], the Monarch integrated ontologies, and other ontologies from Open Biological and Biomedical Ontologies (OBO) [2].
- A continually-updated **knowledge graph and data lake** will provide Translator with a standardized and integrated global view of the whole information landscape.
- **Next-generation Shared Translator Services** will integrate features of ROBOKOP [3], Monarch [4], BioLink [5], and the reasoner APIs to remove integration barriers. These services will provide validation, lookup, and mapping functionality for use across Translator.
- A **registry of Translator KPs, ARAs, and shared services** will increase efficiency, eliminate duplication of effort, and promote collaboration.

2) Our proposed SRI will address the problem of connecting together different components and data/information sources at scale, with community buy-in, and with a plan for sustainability.

3) For the development of the standards component of the SRI, our plan will begin with accepted Translator standards, and we will work with the ARS, ARAs, and KPs to identify gaps. We will have a community process for contributing to the standards, making use of GitHub pull requests and voting, to help everyone contribute effectively and fairly with clear attribution. We will ensure rigorous documentation and testing. For the reference implementation component, we will stand up core Translator services, and will include additional services if they are useful to more than one Translator component rather than used by only one.

4) Consensus-building is hard. Our team has proven expertise and resources to identify needs, refine solutions, and find agreement, thereby successfully bringing infrastructure to fruition. Our team also has the technical and biological expertise to design and test the necessary standards, having been at the forefront of multiple ontology, data standards, and large enterprise software initiatives.

5) The Translator infrastructure is by nature heterogeneous, distributed, and growing; consequently, the most significant data and infrastructure challenge is managing the validity, currency, equivalency, and typing of entities (diseases, phenotypes, drugs, etc.). Our group has developed several innovative algorithms for managing this and related problems; these algorithms are in use for other integration projects and will be modified to suit Translator needs.

Project Plan

The primary goal of the proposed SRI is to enable the ARAs and ARS to leverage the KPs to answer translational questions in a reproducible manner, using a sustainable and collaborative infrastructure. This vision requires a consensus-driven approach to maturing and establishing Translator standards; it also requires normalization services that implement these standards and render the results from one component useful across any other. The feasibility phase of Translator (hereafter "Phase 1") saw a shifting line between federated vs. centralized database approaches, with a trend towards increasing centralization of core knowledge resources. We see this trend continuing toward more core knowledge resources, unified within a wider data lake and supported by federated architecture of analytic tools, reference datasets, and specialized high-volume analytic services.

Our proposal is organized around the following themes:

GC: Community governance coordination. Clear and community-approved governance processes and documentation are required for the Translator to mature and scale sustainably. We view the challenges as being both social and technical: the entire Translator community and broader stakeholder communities will need to be fully engaged and represented by our efforts. To facilitate productive discussions, we plan to establish a preliminary governance strategy during segment 1, drawing upon other experiences and resources, and based on lessons learned during the Translator feasibility phase. We will create a formal consensus building and vetting process, documentation, a library of examples, and a quality assurance pipeline for any relevant components described above. We will draw upon MIT's "A Short Guide to Consensus Building" [6] and the governance guidelines from the scikit-learn [7] project to inform our governance process. We will define and enforce reasonable common coding standards, customized to different requirements to facilitate incorporation of contributions from the community. We will enforce modular development, documentation, release on standard distribution sites (pypi [8], CRAN [9], Maven Central [10], OBO [11], etc.) and the use of testing and continuous integration. Our governance process will enable us to maintain a collaborative, asynchronous, well-documented Translator environment that welcomes contributions from the wider scientific community. We will develop governance rules and SOPs to help everyone contribute effectively and fairly with clear attribution, including third-party "Translator-ready" resources.

AS: Architecture and API specifications. Our group has extensive experience implementing and developing coding standards for highly collaborative, community-driven projects (see Previous Work Examples). We also have demonstrated successes engaging communities without a history of code development. We will implement best practices from industry and scientific computing in project organization and compliance standards (code of conduct, continuous integration checks, versioned public releases, etc.). We will help the program arrive at shared architectural decisions, defining the contracts between KPs, ARAs, and the ARS, and revealing integration and data inclusion barriers. These architectural decisions will also drive further updates to Translator API specifications, such as the ReasonerAPI standard. APIs will be augmented with features such as increased expressivity using logical operations, the ability to provision provenance and attribution information, and versions for all components. Finally, interoperable data licensing continues to be a significant barrier, and we will endeavor to help the KPs improve their own and their sources' licensing terms through the Reusable Data Project (reusabledata.org).

BL: BioLink model. BioLink is already an accepted Translator standard. We will continue to refine this based on a community review of requirements. We will expand the schema to include additional entity types, edge types (predicates), and node and edge properties. We will extend the existing predicate mapping scheme to allow contextual mapping (i.e., selecting different mapped predicates depending on the types of nodes connected), and coordinate the mapping of all predicates from all sources. We will extend the framework to allow modeling the same information from different data modeling perspectives, and define computable transformations between these; for example, allowing a pathway to be represented alternately as a molecule graph vs. a bipartite reaction-chemical graph. We will work closely with the ARAs, ARS, and KPs to develop consensus around a Translator-wide data model.

RO: Integrated reference ontologies. We will provide standard terms, relationships and annotations through a curated set of integrated ontologies based on the Monarch Ontologies and Ubergraph. New development will include curating the set of ontologies to cover all Translator use cases, and inferring cross-ontology linkages. We will integrate ontologies into our services, to take advantage of encoded knowledge to enhance queries. Translator activities will naturally produce improvements to the selected ontologies, which will be fed back to the larger community.

KG: Knowledge graph and data lake. We will provide an up-to-date, BioLink-compliant Knowledge Graph (KG) of all core knowledge resources, plus a virtualized data lake of analytic tools, large datasets, and specialized high-volume analytic services. While ARAs will often directly communicate with a set of federated KPs, a centralized core KG will allow users to perform complex reasoning and analytic queries that would not be possible via chained lookup operations. For instance, standard machine learning approaches such as node and edge embedding would be highly impractical with a distributed approach. At the same time, we recognize the need for distributed access to high-volume or compute resources, and these will be supported as part of a larger virtual Translator data lake, where data and services would be remote, but accessible via a more uniform interface, leveraging the smart API registry. Assembling this KG and data lake will create mapping and alignment problems similar to those that our group has solved for other massive data integration projects by using a suite of innovative algorithms developed in-house. These will be adapted to meet Translator needs.

TS: Next-generation Shared Translator Services. To relieve integration barriers and facilitate interoperability, we will implement an array of functions that support essential integration and knowledge graph operations, primarily focused on common utilities that are needed by multiple projects, or which require a single system-wide implementation. Examples include identifier equivalence methods, name resolution, ontology operations such as semantic similarity, and graph operations such as graph repair, in which rewrites to noncompliant graphs are suggested. We will coordinate APIs currently used in Translator using the best qualities of each to provide a consistent set of tested, maintained, shared services.

TR: A registry of Translator KPs, ARAs, and shared services will increase efficiency, eliminate duplication of effort, and promote collaboration. Resources will be tagged with community-defined, computationally-discoverable metadata, including the results of compliance testing, so that these resources may be automatically discovered and accessed. The registry will also be a central point for generating resource identifiers, and for providing and assigning credit for the publication of resources.

Personnel

Our team includes some of the strongest clinical and translational semantic engineers and community standards developers in the world. As their extensive co-authorship record attests (see Previous Work), our three PIs are experienced leaders, each with a history of driving projects that are collaborative, open, and impactful. Individually and collectively, we have built resources—including standards, ontologies, and tools—that are among the most widely used in biomedical informatics.

Our personnel will be primarily drawn from three institutions: **Oregon State University (OSU), Lawrence Berkeley National Laboratory (LBNL),** and the **Renaissance Computing Institute at the University of North Carolina (RENCI).** All of these have a strong track record in open source bioinformatics. These lead institutions, in turn, will draw upon the needed expertise provided by some of our close collaborators from several additional organizations. All of the people on our team have worked together on successful multi-organization projects (including the first phase of Translator), as listed in the "Previous Work Examples" section.

Multi-PIs

Dr. Melissa Haendel is the Director of Translational Data Science at OSU, and directs the Translational and Integrative Science Lab (TISLab; tislab.org), a cross-disciplinary team of biologists, bioinformaticians, clinicians, and software developers with expertise in biomedical ontology development, semantic data modeling, data integration pipelines, and graph database construction. The group is also known as a leader in innovative program management of large consortial groups, and teaches courses on technical program management. Dr. Haendel co-leads the Monarch Initiative (with Dr. Mungall), a cross-species genotype-to-phenotype discovery consortium, and was also a multi-PI for the Orange team in the first phase of the NCATS Translator project (with Dr. Mungall). She has demonstrated success in leadership of cross-disciplinary international teams, development of applications used for rare disease diagnostics, implementation of platforms and tools for translational research, and open and reproducible science. She was a member of the Biden Cancer Moonshot Blue Ribbon Panel on Data Sharing, is on the NCI Semantic Scientific Committee, and leads the Clinical and Phenotypic workstream within the GA4GH, an international organization developing standards for the transfer of genomic knowledge.

Dr. Christopher Mungall is the head of LBNL's Molecular Ecosystems Biology department. For two decades, Dr. Mungall has been a recognized leader in the development and use of biomedical ontologies and knowledgebases to advance scientific research. He is a leading expert in the semantic modeling of biomedical data. He and his team have developed ontologies that encode formal, computable definitions for gene function, anatomy, phenotypes, human diseases, and environmental states. Dr. Mungall is a co-PI for the Gene Ontology (GO), the most widely used ontology in the biological research community. He is also one of the founders and leaders of the Open Biomedical Ontologies (OBO) Foundry, and a co-PI of the Monarch Initiative (with Dr. Haendel). He was a multi-PI for the Orange team in Translator Phase 1 (alongside Dr. Haendel), where his team worked on the BioLink model and multiple Translator components, such as Knowledge Beacons. In addition to GO, Dr. Mungall leads or co-leads the development of a number of key community-driven biological ontologies including Uberon, the Cell Ontology, uPheno, and Mondo. The areas of expertise provided by Dr. Mungall and his group include artificial intelligence, data modeling, knowledge graphs, ontology engineering, and defining biomedical standards. Beyond Dr. Mungall's group, the team can draw upon the

extensive expertise of other scientists and engineers across LBNL, as well as the computing and networking infrastructure available to support research endeavors.

Dr. Chris Bizon is the Director of Analytics and Data Science at RENCI**,** where he leads a group of nine scientists and developers working on numerous data science projects. He is the co-PI on an NCATS Data Reasoner OTA, leading the development of the ROBOKOP tool, which integrates a wide range of biomedical knowledge and allows graph-based queries providing novel insights. He also co-leads ClinGen's Data Model Working Group. This cross-grant working group creates data models and exchange formats to allow unambiguous communication about genetic variants and the structured interpretations of their pathogenicity. This work has led to related efforts performed under the auspices of the GA4GH. Dr. Bizon and his group are proficient at creating enterprise software that provides useful and scalable services. They have extensive experience developing ontology-based software applications and implementing automated reasoning within these applications.

Additional Key Personnel

Dr. James Balhoff (RENCI) is a Senior Research Scientist at the Renaissance Computing Institute (RENCI) at the University of North Carolina at Chapel Hill. Dr. Balhoff has expertise in the application of ontologies and other semantic technologies to facilitate reuse and integration of biological data. He has developed a variety of semantic tools supporting shared infrastructure across multiple large projects including Gene Ontology, Monarch Initiative, NCATS Translator, and the NSF-funded Phenoscape Project. In the first phase of Translator he developed the ROBOKOP Ubergraph.

Dr. Matt Brush (OHSU) will serve as the site-PI for OHSU. Dr. Brush is the Lead Ontologist and Research Assistant Professor for the Translational and Integrative Sciences Lab (TISLab) in the Department of Medical Informatics and Clinical Epidemiology (DMICE). He is an expert in semantic requirements analysis, data model landscape analysis, cancer evidence and provenance representation and harmonizing data model with genomics standards from GA4GH. During the first phase of Translator, Dr. Brush worked on the BioLink model and led predicate mapping efforts.

Kent Shefchek (OSU) is the lead, full stack developer of the Monarch Initiative. He develops software for data ingest, integration, and publication. He is responsible for the primary development of Monarch data integration and services technologies via SciGraph, and has been involved in developing the API that serves phenomic and disease information to the research communities. He has also contributed significantly to Monarch's API, which was the first instantiation of a BioLink API.

Dr. Anne Thessen (OSU) is the lead on the GenoPhenoEnvo project funded by the National Science Foundation *Harnessing the Data Revolution* program. She is also a semantic engineer for the *Center for Cancer Data Harmonization* project funded by NCI and co-leads the NHLBI BioData Catalyst project data harmonization working group. She is the lead developer of ECTO and ECOCORE ontologies and a contributor to the Ontology for Biological Attributes and the Environment Ontology. She is also heavily involved in standards development through the Research Data Alliance. Before joining TISLab, she operated her own data science consulting company for five years, during which time she worked on the Encyclopedia of Life TraitBank, the Deepwater Horizon Database, and PollardBase.

Dr. Patrick Wang is a Research Engineer at CoVar Applied Technologies and an Adjunct Assistant Professor of Electrical and Computer Engineering at Duke University. He helped

develop the Translator feasibility-phase API standards and was a key contributor to the ROBOKOP project, serving as lead developer of the service architecture, question-parsing and ranking components, which gave him a deep familiarity with the requirements and challenges involved in connecting the varied services of the Translator ecosystem. Dr. Wang's substantial experience building enterprise software across biomedical, defense, and commercial applications will be an asset in engineering robust "reference implementations" of shared Translator tools and infrastructure.

Additional Personnel

Dr. Stan Ahalt (RENCI) is the director of RENCI and the associate director of Informatics and Data Science at the NC TraCS Institute, as well as a Professor of Computer Science at the University of North Carolina, Chapel Hill, and the head of the Steering Committee for the National Consortium for Data Science. Dr. Ahalt was a PI on the earlier feasibility phase of Translator for the Green team.

Steven Cox (RENCI) specializes in developing enterprise software for large, distributed projects. He is the Cyberinfrastructure Engagement Lead at RENCI, where he designs scalable data science platforms for data-intensive challenges in infrastructure instrumentation, machine learning, visualization, and the integration of disparate biomedical data stores. Mr. Cox has played a leading role in the development and deployment of cloud services in RENCI's NCATS Data Translator and NHLBI Data STAGE projects, gaining expertise in platforms including Gen3.

Nomi Harris (LBNL) has a long history in bioinformatics, starting as a software developer and transitioning into scientific project management and communication. Her leadership in the bioinformatics open source community includes chairing the annual Bioinformatics Open Source Conference (BOSC) for the past nine years and serving as a board member of the Open Bioinformatics Foundation. As the program manager for the Mungall group, Ms. Harris works to coordinate an extensive portfolio of large multi-institution bioinformatics-related projects and ensure smooth operation and clear communication between collaborators and with the public. She has worked on the NCATS Translator, Gene Ontology, Monarch Initiative, and other multi-institution projects.

Dr. Maureen Hoatlin (OHSU) is an Associate Professor of Biochemistry and Molecular Biology at OHSU. Her research involves discovery and analysis of novel proteins that control cancer susceptibility, and targets for novel cancer therapeutics, with a special focus on the Fanconi anemia/Breast Cancer protein network where defects lead to bone marrow failure, hematologic malignancy and other cancers. She has experience in the biotechnology industry and academic settings; a recent MBA and drug development pipeline training; and a history of success on multi-site matrixed collaborations. Dr. Hoatlin led the Fanconi anemia demonstrator, which defined competency questions, defined workflows, and analyzed and compared results. The Fanconi demonstrator was utilized by many Translator teams in the feasibility phase under Maureen's leadership and guidance.

TBN (OSU) will serve as the project manager for the SRI. This individual will have previous experience as a project manager for geographically distributed scientific software projects. They will also have experience in online community management, governance coordination, and outreach.

Deepak Unni (LBNL) has extensive experience in genomics, bioinformatics and software engineering. He has worked on large collaborative projects such as the Generic Model Organism Database (GMOD) project, as well as being a key developer on the Apollo collaborative genome annotation tool and the JBrowse genome visualization application. Mr. Unni has worked on the Monarch Initiative project developing the BioLink API and the underlying database system. He also played a key role in the first phase of Translator designing and implementing APIs and the BioLink model, and he developed the Knowledge Graph Exchange (KGX) toolkit.

Contractors

Dr. Richard Bruskiewich is the Founder and CEO of the STAR Informatics Group / Delphinai Corporation, a consulting firm specializing in developing cutting edge scientific software and database systems for diverse biomedical and scientific applications. He is also an Adjunct Professor in the Department of Botany at the University of British Columbia (UBC). Dr. Bruskiewich was involved in the first phase of Translator, developing the Knowledge Beacon[12] system and implementing Translator workflows.

Kenneth Huellas-Bruskiewicz has worked on diverse software development projects for STAR for three years. He participated in the first phase of Translator, converting Jupyter Notebooks into modules usable in a workflow management system such as CWL and developing a working prototype of the “Gene List Sharpener” web interface.

Dr. Karamarie Fecho is a RENCI-affiliated biomedical consultant with over 20 years of experience spanning the clinical and translational spectrum and focused on diverse domains of significance to the proposed work, including neurobiology, psychoneuroimmunology, anesthesiology, and pharmacology. She is currently an investigator on two Translator proposals, leading the design, development, and evaluation of the prototype ICEES and working closely with institutional regulatory bodies to ensure that we maintain compliance as we develop capabilities to openly expose integrated clinical and environmental data. She has also contributed to the design and testing of TranQL and ROBOKOP.

Harold Solbrig is an Assistant Professor at The Johns Hopkins University School of Medicine. He has worked for 30 years in terminologies, ontologies and semantic data modeling, and is the primary author of the HL7 Common Terminology Services 2 (CTS2) specification and a major contributor and editor on part 3 of the ISO/IEC 11179-3 Metadata Registries standard. He is a co-developer of the Shape Expressions (ShEx) language and the author of the Python ShEx implementation. He is a key contributor to the FHIR RDF specification and implementation and has been actively involved in the BioLink Modeling Language (BioLinkML) toolkit. He was a key contributor during the first phase of Translator.

Table of roles and responsibilities

Organizing themes of the proposed work:

- GC Community governance coordination
- AS Architecture and API specifications
- BL BioLink model
- RO Integrated reference ontologies
- KG Knowledge graph and data lake
- TS Next-generation shared Translator Services
- TR A registry of Translator KPs, ARAs, and shared services

Oregon State University/Oregon Health & Science University (TISLab)								
Person	**Role**	GC	AS	BL	RO	KG	TS	TR
Melissa Haendel	MPI; translation vision. Governance and community coordination; BioLink model and integrated ontology development; translational demonstrators; overseeing biological vision/validating modeling/KGs	x		x	x			
Kent Shefchek	Database and API engineering; architecture, unit test development		x				x	
Anne Thessen	Ontology development, BioLink development; registry development; architecture working group		x	x	x			x
TBN Project Manager	Community management; governance coordination and implementation; outreach; documentation	x	x	x	x	x	x	x
Harold Solbrig	BioLink-model-FHIR clinical integration; BioLink model framework			x				
Maureen Hoatlin	SME; requirements and landscape analysis; community coordination	x	x					
Matt Brush	BioLink model data modeling; ontology development and engineering; evidence and attribution modeling			x	x			
Lawrence Berkeley National Laboratory								
Person	**Role**	GC	AS	BL	RO	KG	TS	TR
Chris Mungall	MPI; integration vision. Governance of standards, shared data modeling, and ontologies; overseeing implementation of modeling and knowledge graphs	x	x	x	x	x		
Deepak Unni	Development of service APIs and implementation framework for BioLink and KGX toolkit; building test frameworks and leading QC pipeline development; coordination of ETL efforts across Translator		x	x			x	
Nomi Harris	Program management and communications	x	x	x	x	x	x	x
Richard Bruskiewich	Implementation and standards development; data modeling in BioLink		x	x			x	
Kenneth Huellas-Bruskiewicz	Implementation and standards development		x	x			x	
University of North Carolina RENCI								
Person	**Role**	GC	AS	BL	RO	KG	TS	TR
Chris Bizon	MPI; enterprise vision. Governance, strategy, and oversight of architecture and API standards; implementation of shared infrastructure and translator registry	x	x				x	x
Jim Balhoff	Graph database design and maintenance; reference ontology integration; kBOOM implementation				x	x		
Steve Cox	Architecture and APIs, with a focus on scalable query and API usability		x				x	x
Karamarie Fecho	Clinical BioLink modeling; interaction with the clinical components; communications and user engagement	x	x	x				
Stan Ahalt	Communications, outreach, and publications	x						
Patrick Wang	Implementation of shared services and translator registry components; development of API standards and overall architecture		x				x	x

Resources

Oregon State University (OSU) and Oregon Health and Sciences University (OHSU).

Due to the nature of Dr. Haendel's shared OSU/OHSU appointment and lab (tislab.org) the proposed work is able to benefit from complementary facilities at each of these sites.

OSU. Although the primary compute infrastructure for this program is currently located in the Amazon cloud, this is complemented by substantial infrastructure within the Center for Genome Research and Biocomputing (CGRB) at OSU. The CGRB's biocomputing infrastructure includes a distributed service architecture, a large compute cluster, and a secure network, supported by seven full-time staff and one part-time statistician dedicated to biocomputing and bioinformatics. Services provided by the CGRB include: provision of data management software including tools for analysis and distribution of high throughput DNA sequence (HTS) and biological imaging data; setup, optimization, maintenance and backup of the research computers owned by investigators and housed in the CGRB server room; access to grid computing resources on the CGRB cluster or on external clusters; access to virtual servers with web and database services; programming instruction, and use of bioinformatics software; hosting a large variety of in-house and commercial bioinformatics tools; statistical consultation, experimental design and analysis.

OHSU. The Department of Medical Informatics and Clinical Epidemiology (DMICE) at OHSU is a rich intellectual environment of more than 80 faculty members, internationally recognized for their accomplishments and innovation in bioinformatics and computational biomedicine.

Lawrence Berkeley National Laboratory (LBNL)

LBNL conducts research across a wide range of scientific disciplines, with key efforts in fundamental studies of the universe, quantitative biology, nanoscience, new energy systems and environmental solutions, as well as the use of integrated computing as a tool for discovery. Thirteen Nobel prizes are associated with LBNL. Fifty-seven Lab scientists are members of the National Academy of Sciences (NAS), one of the highest honors for a scientist in the United States. Thirteen of our scientists have won the National Medal of Science, 18 of our engineers have been elected to the National Academy of Engineering, and three of our scientists have been elected into the Institute of Medicine. LBNL has been a leader in creating the modern-day model of team science. This is exemplified by the numerous national user facilities operated by the Lab, including the Advanced Light Source, Energy Science network (ESnet), the Joint Genome Institute (JGI), the National Energy Research Scientific Computing Center (NERSC), and the Lawrencium cluster. These facilities are available to the project team as needed.

Renaissance Computing Institute (RENCI)

An institute of the University of North Carolina at Chapel Hill, The Renaissance Computing Institute has extensive expertise in scientific software development. Since 2004, RENCI has served as a living laboratory fostering data science expertise, advancing software development tools and techniques, developing effective cross-disciplinary and cross-sector engagement strategies, and establishing sustainable business models for software and services. RENCI's cyberinfrastructure is equipped to support a variety of projects and activities, including those proposed as part of the Translator SRI. The available resources include the 2000-square-foot Europa Data Center, which hosts a range of high-performance computational infrastructure, as well as cluster storage and fast network infrastructure.

Previous Work Examples

Community Resource	Description	Partici-pants
BioLink [5]	The BioLink model is an upper-level data model for representing biological knowledge. It was launched by our group in the feasibility phase of Translator to harmonize knowledge graphs (e.g., semantic types and predicates).	All
GO [13,14]	The Gene Ontology knowledgebase is the world's largest source of information on the functions of genes. The mission of the GO is to develop a comprehensive computational model of biological systems, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life.	CM, JB, NH
OBO [11]	Open Biomedical Ontologies (OBO) is a community effort to create controlled vocabularies for shared use across different biological and medical domains. The mission of the OBO Foundry is to develop a family of interoperable ontologies that are both logically well-formed and scientifically accurate.	CM, MH, JB, MB
Monarch [4,15]	The Monarch Initiative is an international consortium dedicated to the integration of model organism and human genotype-phenotype data using semantic technologies.	MH, CM, JB, MB, DU, KS, AT, NH
Human Phenotype Ontology (HPO) [16]	The HPO, one of the most widely used biomedical ontologies and the gold standard for describing human phenotypes, is an IRDIRC-recommended resource, and is used by thousands of organizations and tools, such as the NIH (NHGRI, NIAID, UDP, etc.), Genomics England, and ClinGen.	CM, MH
Pheno-packets [17]	Phenopackets, a standard format for sharing phenotypic information, was developed in collaboration with GA4GH; they have newly approved Phenopackets as a GA4GH standard [18]. It aims to enable an ecosystem of interoperable tools and resources for sharing and utilizing case-level phenotypic data.	CM, MH, KS
ODK [19]	The Ontology Development Kit packages all aspects of the ontology release cycle in a single, simple to use framework, and provides a widely used standard for ontology releases.	CM, JB
NMDC [20]	The National Microbiome Data Collaborative (NMDC) is developing an open-access framework that facilitates more efficient use of microbiome data for applications in energy, environment, health, and agriculture.	CM, DU
Mondo integrated disease ontology [21]	Mondo is a unified disease ontology that harmonizes disease definitions by integrating disease terms from multiple resources to yield a logically coherent merged ontology. It is used by Monarch and has been adopted by ClinGen for their disease-gene associations, by	CM, MH, MB, JB, DU

	GARD, by commercial entities such as SunQuest, and by the EBI as the basis for EFO3's disease content.	
ROBOKOP [3]	ROBOKOP (Reasoning Over Biomedical Objects linked in Knowledge-Oriented Pathways) is a biomedical reasoning system, developed in Translator Phase 1, that interacts with many biomedical knowledge sources to answer questions.	CB, PW, JB, SC, KF
ICEES [22]	ICEES, created during Translator Phase 1, is a pipeline and interface that publicly exposes joined clinical and environmental data, meeting all regulatory compliances.	KF, SC, SA, CB
Reasoner API [23]	The ReasonerAPI standard was developed to provide a common method for posting questions and receiving answers from Reasoning tools during the Translator feasibility phase.	PW, SC
TranQL [24]	TranQL, developed during Translator Phase 1, is a query/workflow language for interactive exploration of federated knowledge graphs.	SC, KF, CB, PW
ClinGen Interpretation Model [25]	The ClinGen Interpretation Model provides a structure and format for exchanging information about the clinical effect of genetic variants. It leverages the Monarch Initiative's Scientific Evidence and Provenance Information Ontology (SEPIO).	CB, MB
GA4GH Variant Annotation [26]	The Variant Annotation data standard is being developed by the GA4GH Genomic Knowledge Standards (GKS) Workstream. It provides a unified modeling framework to represent diverse kinds of knowledge annotated to genetic variation, and enables its exchange and application across knowledge curation and clinical decision support systems.	CB, MB, MH
Knowledge Beacons [27]	The Knowledge Beacon REST API (developed during Translator Phase 1) is a BioLink-compliant specification for a web service that retrieves concepts, relationships between concepts, and associated provenance, in a uniform manner, from online knowledge sources exposing the API.	RB, KHB, CM
GMOD [28]	Generic Model Organism Database (GMOD) provides research communities with a toolkit of open-source software components for visualizing, annotating, managing, and storing biological data.	CM, NH
Uberon [29], CL [30], uPheno [31]	Ontologies that team members have developed include: • Unified Phenotype Ontology (uPheno): cross-species phenotypes • Cell Ontology (CL): cell types in animals • Uberon: anatomical structures in animals	CM, MH, MB, AT
CD2H [32]	In the National Center for Data to Health (CD2H), informatics experts across 60 institutions are developing standardized approaches and best practices; algorithms such as LOINC2HPO, communities of practice such as the Health Open Terminology (HOT) ecosystem, and FHIR-based data harmonization resources to address operational and barriers to sharing and using clinical and translational data.	MH, HS

Data Sharing and Collaboration Plan

Overview. Our SRI is comprised of multiple open-source, open-access components: data, ontologies and other standards, software tools, and algorithms. **We want researchers to be able to easily discover, search, explore, and download all of Translator's work.**

Open science has been at the heart of our highly collaborative work since its inception. We will continue to provide community access to all the resources we generate, links to the source code, and documentation that will help improve the accessibility and reusability of our products. Software, data, and other resources will be released with licenses that enable further reuse, modification, and redistribution by downstream users and developers.

The Previous Work section above lists the Translator-relevant components that we have built to date; the referenced bibliography therein points to corresponding **documentation and open source code in GitHub.**

- **Data.** Our sharing plan adheres to NIH Grant Policy [33]. Throughout the development process, key releases of data will be deposited in Zenodo, generating an associated DOI and listing ORCIDs for contributors. We will version all constituent data sources as well as integrated ones, and promote similar best practices for KPs. We will evaluate our data sharing according to the rigorous rubrics that we worked to develop [34–36], especially licensing and identifier hygiene.
- **Ontologies and vocabularies.** We will use standard open vocabularies wherever possible in the process of harmonizing and integrating data for the SRI. As we are the developers of many of these open vocabularies (see Previous Work) we are well positioned to both leverage and extend them where necessary for this proposal.
- **Data models and exchange standards.** We will refine and implement open data exchange standards. The standards themselves (see Previous Work) will remain open access, open source, community-driven, and permissibly licensed.
- **Software.** We always make a **standard and stable version of each piece of code available** for public download, with the source code stored in GitHub. Where practicable and useful, we will also continue to make our software accessible via REST APIs. All software written for this project will be released under the most liberal license possible given institutional limitations, with the default being the BSD 3-Clause License [37].
- **Attribution.** We will help the Translator community register their resources in public registries, such as the CD2H/CTSA Tool Registry. We will explore and help develop the newly created Contributor Attribution Model (CAM) [38] and associated Contributor Role Ontology [39] for attribution of all artifacts.

Collaboration. All of the aforementioned best practices support collaboration. Use of GitHub issue trackers enables community members to submit questions, bug reports and suggestions that will be visible not only to project members but to other users as well, consistent with our commitment to transparency and community development. We will encourage those who fork, modify or add to our code to submit it to GitHub to be considered for inclusion into our overall codebase. We will support office hours in the context of the SRI, and asynchronous communication via Slack. We also endeavor to have informational sites for key components (for example, the in-development one for the BioLink model), and will help standardize how this information is collected and presented within the Translator.

Milestones for all segments, years and themes

Milestone	Description	Timing
GC. Community governance coordination will be developed with community buy-in to ensure an effective collaborative environment, and drive consortium-wide consensus on the other components.		
GC.1: Strawman governance docs	Draft procedures for decision-making and document how these procedures themselves will be updated as necessary over time.	Y1a
GC.2: Governance documentation	Documents decision-making processes and SOPs for proposing new standards, improving existing standards, implementing new processes, and any additional needed community decision-making.	Y1a,b, c Y2-Y5
GC.3: Standards and architecture decision-making	Execute a standard set of processes for each component as it comes to fruition for community vetting and release/endorsement. For example, the SRI team will work closely with KP, ARA, and KRS teams to design the operational contract defining which parties are, in general, responsible for various transformations and semantic operations. These kinds of decisions are key to creating reliably interoperable and reusable components and the SRI will aid documentation and dissemination of such decisions.	Y1b-Y 5
GC.4: Collaboration support	Create virtual opportunities to facilitate collaboration, such as pair-programming, mentoring, and office hours.	Y1b-Y 5
GC.5: Conflict resolution	Establish conflict resolution strategies and a code of conduct.	Y1a
GC.6: Sustainability plan	Develop and refine a plan to ensure the continued community-driven evolution of Translator beyond the 5-year project period. We will nurture an external developer community and create products that fill their needs. For contributed services we will follow a similar model of community contribution, but will enforce common coding standards, including modular development, documentation, release on standard distribution sites (pypi, CRAN, Maven Central, OBO, etc.) and the use of testing and continuous integration.	Y1b Y1c-Y5
AS. Architecture and API specifications will drive community efforts to define details of project architecture and communication protocols across KPs, ARAs, and ARS.		
AS.1: Documentation of shared implementation protocols	The documentation will include: ● Guidelines for federation vs. integration to help Translator teams efficiently divide labor ● A semantic model for Translator workflow operations to maintain provenance of a query by providing a controlled vocabulary to characterize workflow operations ● Language of discourse for user-facing API for Translator operations ● Identifier best-practices based on prior work and Translator need ● API standards that define the interfaces used for interoperability between KPs, ARAs, the ARS, and other services ● Protocols for handling user queries ● Process for handling user feedback (See also Governance above)	Y1b Y2-5

AS.2: Validation and test suites	Develop validation and test suites for all KPs and ARAs to ensure delivery of accurate and repeatable results with appropriate provenance.	Y2-Y5
AS.3: Data QC and versioning standards	Coordinate a consensus protocol for long-term data quality control and data updates. In addition to Translator-wide standardization, the API will include QA/QC services. There will be a weekly/monthly/daily quality check of data and tests of services using standard protocols, and a defined versioning system.	Y1-5
AS.4: Develop metadata standards	Create computationally actionable metadata for all Translator products, including data sets, tools, and APIs to support versioning and provenance.	Y1-5
AS.5: Provide documentation portal	Develop a documentation website and structure to allow KP/ARA owners to easily add their own documentation.	Y1
AS.6 Workflow standardization coordination	We will assist the ARAs, KPs, and ARS in standardizing and automating workflows. This will include standards and SOPs for knowledge retrieval, transformation, workflow iteration, and assessment/combination of evidence. (note that an ARA for the technical aspects of this work is also being submitted)	Y2-Y5
BL. The BioLink model will define the standard entity types, relationship types, and a schema shared by all Translator components.		
BL.1: BioLink community review and requirements	Review the BioLink model with the APAs and KPs to ensure that it meets the needs of Translator. Make plans for the addition of any missing components.	Y1b-Y5
BL.2: Understand landscape of perspectives	Collaborate with the KPs and ARAs to form a list of perspectives to represent in BioLink. Plan implementation or accommodation of these perspectives.	Y2-Y3
BL.3: Predicate mapping framework and implementation	Create framework for predicate mapping that extends the existing pairwise mappings, to allow for contextual mapping (e.g., if the predicate is 'increases', and this connects a gene to a disease, map to 'increases risk of'). Curate mappings for the 10,000+ source predicates gathered in the first phase of Translator using a combination of consortium-wide crowdsourcing and expert curation. https://github.com/biolink/biolink-model/issues/285; https://github.com/biolink/biolink-model/issues/286	Y1a-Y5
BL.4: Ongoing curation of model	Continually adapt and extend model in response to needs of Translator consortium. This includes improving documentation, adding or modifying mappings, and adding constraints.	Y1a-Y4
BL.5: Implement views	Implement alternate views/perspectives in the model, and define computable transformations between them; e.g., between interaction-centric views and pathway-centric views. See https://github.com/biolink/biolink-model/issues/284	Y1a-Y3
BL.6: Validation and reporting module	Extend and harden the existing validation suite to include checks of (a) consistency between node types and ontology descriptors; (b) declaration of node edge labels (predicates); (c) definitions of properties assigned to nodes or edges; (d) consistency between predicates and domain, range, and other constraints.	Y1a-Y5
BL.7: Integrate logical inference into	Extend the validation framework with logical inference using biological knowledge encoded in existing ontologies. We will extend this with a general framework that	Y2-Y5

validation framework	will allow us to specify rules for different domains, for example, to ensure that reactions are balanced. See https://github.com/biolink/biolink-model/issues/287	
BL.8: Modularize, extend, and harden KGX toolkit	The KGX toolkit can be used to mix and merge subsets of existing knowledge graphs, performing clique merges on equivalence sets. We will modularize this and offer it as a service.	Y2-Y5
RO. Integrated reference ontologies will provide BioLink-compliant terms and relationships.		
RO.1: Prototype semantic type inference service	This prototype API will take a CURIE and return a list of BioLink model types for the CURIE.	Y1a
RO.2: Semantic type inference service	Harden and maintain the prototype created in Segment 1 with unit tests, extended BioLink model, and type-specific services.	Y2-Y5
RO.3: Review set of ontologies required for Translator	Gather the set of ontologies currently used in Phase 1, plus those added in Segment 1, and review and create recommendations. Where there are duplicative ontologies in a domain, we will either recommend one, or propose a way to combine them into a non-redundant set.	Y1b-Y2
RO.4: Translator integrated ontology graph	Create a pipeline to take the combined set of Translator ontologies and merge these together, implementing logical reasoner-based consistency checks. The end product will be versioned, alongside the source versions.	Y2-Y3
RO.5: Maintain integrated ontologies	Maintain the curated set of ontologies over time, plus creating any necessary bridge axioms, feeding back to source ontologies where appropriate.	Y2-Y5
KG. A continually-updated **knowledge graph and data lake** will provide Translator with a standardized and integrated global view of the whole information landscape.		
KG.1: Deliver Translator knowledge graph endpoint	The core KG will make use of centralized ETL pipelines to make it easy for groups to contribute their modules and content without duplicating effort. We will extend our existing clique merging techniques and our probabilistic kBOOM algorithm to merge equivalent entities that use different identifiers from different sources. Provenance of results will be maintained using SEPIO and standardized semantic workflows.	Y1b-Y5
KG.2: Harden processes for KG	Optimize the ETL and kBOOM modules for Translator; harden processes that create the KG endpoint	Y2
KG.3: Develop QA/QC and consistency protocols	In order for users to trust the knowledge graph and data lake, we need to provide basic reports on the quality and consistency of the data after every build. This should include versioning.	Y1b-Y2
TS. Next-generation Shared Translator Services will integrate features of ROBOKOP, Monarch, BioLink, and reasoners to remove integration barriers.		
TS.1: Prototype equivalence service for disease and phenotype entities	Build a prototype API that takes as input a CURIE from any disease of phenotype ontology and returns all known equivalent or closest matching CURIES. (e.g., Mondo, SNOMED, ICD-9,10,11; ICD-O, DO, MeSH, UMLS, HP, Model Organism phenotypes).	Y1a
TS.2: Prototype equivalence service	This prototype API will take as input a CURIE from any chemical ontology/vocabulary and return all known equivalent or closest matching CURIES.	Y1a

for chemical entities	(e.g., ChEBI, CHEMBL, DRUGBANK, PUBCHEM, MESH, HMDB, INCHIKEY, UNII, KEGG.COMPOUND, GTOPDB)	
TS.3: Prototype compliance checking for KP	This prototype API will check compliance with BioLink for the KPs. If a resource is not compliant, the API will return an informative message about the specific failure. This will be developed in collaboration with the KPs.	Y1a
TS.4: Prototype predicate translation service	Build a prototype API endpoint that returns the 'translator' specification for a predicate given as input.	Y1a
TS.5: Prototype name resolution services	Build a prototype API endpoint that takes a text string and return CURIES for entities that match using NLP matching over named entities in the Translator knowledge graph.	Y1a
TS.6: Harden prototypes from Segment 1a	Based on the prototyping experience and results, refactor the segment 1a prototypes into fully operational and deployed components on shared infrastructure.	Y1b
TS.7: Extend equivalence API to handle complex path matches and semantic similarity	The segment 1 equivalence API will return equivalent or exact match CURIEs. In segment 2 we will return entities that have more complex relationships. We will extend the equivalence API and return the full logical structure or graph that captures the relationship. We will also implement generic semantic similarity.	Y1b-Y5
TS.8: Object registration service	Develop a global Translator object registration service that can assign unique, standardized, and persistent identifiers to data objects. These identifiers can be used across the Translator project for data discovery and provenance.	Y2-Y5
TS.9: New service development	Based on integration or development barriers discussed through the architecture component, implement new shared services that alleviate issues or provide a common means to arrive at shared results.	Y2-Y5
TS.10: User and developer docs	Create documentation on how to use and develop for the Translator APIs.	Y1b-Y5
TR. A **registry of Translator KPs, ARAs, and shared services** will increase efficiency, eliminate duplication of effort, and promote collaboration.		
TR.1: Implement registry	Develop a registry consisting of all existing KPs, ARAs, and other shared services for reference by all Translator projects.	Y1b-Y5
TR.2: Discovery interfaces	Provide APIs to the registry allowing computational discovery of relevant tools and KPs.	Y2-Y5
TR.3: DOIs and micropublications	Facilitate the provisioning of DOIs and micropublications for registry entries, using the Contributor Attribution Model.	Y3-Y5
TR.4: Integrate with compliance testing	Provide automatically generated per-entry compliance and uptime information.	Y2
TR.5: Maintain and refine registry	Maintain and update the registry as needed, and provide assistance to registrants in its use.	Y1b-Y5

BIBLIOGRAPHY

1. ubergraph. Github; Available: https://github.com/NCATS-Tangerine/ubergraph
2. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25: 1251–1255.
3. ROBOKOP. [cited 14 Nov 2019]. Available: https://ROBOKOP.renci.org/
4. Mungall CJ, McMurry JA, Köhler S, Balhoff JP, Borromeo C, Brush M, et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res. 2017;45: D712–D722.
5. biolink-model. Github; Available: https://github.com/biolink/biolink-model
6. A Short Guide to Consensus Building | publicdisputes. [cited 14 Nov 2019]. Available: https://publicdisputes.mit.edu/short-guide-consensus-building
7. Scikit-learn governance and decision-making — scikit-learn 0.21.3 documentation. [cited 13 Nov 2019]. Available: https://scikit-learn.org/stable/governance.html
8. PyPI · The Python Package Index. In: PyPI [Internet]. [cited 14 Nov 2019]. Available: https://pypi.org/
9. The Comprehensive R Archive Network. [cited 14 Nov 2019]. Available: https://cran.r-project.org/
10. Maven Repository: Search/Browse/Explore. [cited 14 Nov 2019]. Available: https://mvnrepository.com/
11. The OBO Foundry. [cited 14 Nov 2019]. Available: http://www.obofoundry.org/
12. translator-knowledge-beacon. Github; Available: https://github.com/NCATS-Tangerine/translator-knowledge-beacon
13. Gene Ontology Resource. In: Gene Ontology Resource [Internet]. [cited 14 Nov 2019]. Available: http://geneontology.org/
14. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019;47: D330–D338.
15. Welcome to Monarch. [cited 14 Nov 2019]. Available: https://monarchinitiative.org/
16. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Res. 2017;45: D865–D876.
17. Phenopackets. Github; Available: https://github.com/phenopackets
18. Phenopackets: Standardizing and Exchanging Patient Phenotypic Data. [cited 14 Nov 2019]. Available: https://www.ga4gh.org/news/phenopackets-standardizing-and-exchanging-patient-phenotypic-data/
19. ontology-development-kit. Github; Available:

https://github.com/INCATools/ontology-development-kit

20. nlopez. National Microbiome Data Collaborative. In: National Microbiome Data Collaborative [Internet]. [cited 14 Nov 2019]. Available: https://microbiomedata.org/
21. mondo. Github; Available: https://github.com/monarch-initiative/mondo
22. ICEES. [cited 14 Nov 2019]. Available: https://researchsoftwareinstitute.github.io/data-translator/apps/icees
23. NCATS-ReasonerStdAPI. Github; Available: https://github.com/NCATS-Tangerine/NCATS-ReasonerStdAPI
24. tranql. Github; Available: https://github.com/NCATS-Tangerine/tranql
25. Interpretation Model. [cited 14 Nov 2019]. Available: https://dataexchange.clinicalgenome.org/interpretation/
26. [No title]. [cited 14 Nov 2019]. Available: https://ga4gh-gks.github.io/variant_annotation.html
27. translator-knowledge-beacon. Github; Available: https://github.com/NCATS-Tangerine/translator-knowledge-beacon
28. GMOD. [cited 14 Nov 2019]. Available: http://gmod.org/wiki/Main_Page
29. Mungall C. Uberon. Uberon Website. [cited 14 Nov 2019]. Available: https://uberon.github.io/
30. Wg OT. Cell Ontology. [cited 14 Nov 2019]. Available: http://www.obofoundry.org/ontology/cl.html
31. upheno. Github; Available: https://github.com/obophenotype/upheno
32. CD2H (Data2Health). Github; Available: https://github.com/data2health
33. of Health NI, Others. NIH data sharing information: Main page. 2007. Available: https://grants.nih.gov/grants/policy/data_sharing/
34. Haendel M, Su A, McMurry J. FAIR-TLC: Metrics to Assess Value of Biomedical Digital Repositories: Response to RFI NOT-OD-16-133. 2016. doi:10.5281/zenodo.203295
35. Carbon S, Champieux R, McMurry JA, Winfree L, Wyatt LR, Haendel MA. An analysis and metric of reusable data licensing practices for biomedical resources. PLoS One. 2019;14: e0213090.
36. (Re)usable Data Project. [cited 23 Oct 2019]. Available: http://reusabledata.org
37. The 3-Clause BSD License | Open Source Initiative. [cited 14 Nov 2019]. Available: http://opensource.org/licenses/BSD-3-Clause
38. Welcome to the Contributor Attribution Model — Contributor Attribution Model documentation. [cited 14 Nov 2019]. Available: https://contributor-attribution-model.readthedocs.io/en/latest/
39. Wg OT. Contributor Role Ontology. [cited 14 Nov 2019]. Available: http://www.obofoundry.org/ontology/cro.html